Genomics 100 (2012) 14–17

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

CACG: A database for comparative analysis of conjoined genes

Dae-Soo Kim [a,1], Dong-Wook Kim [a,1], Min-Young Kim [a], Seong-Hyeuk Nam [a], Sang-Haeng Choi [a], Ryong Nam Kim [a], Aram Kang [a,b], Aeri Kim [a,b], Hong-Seog Park [a,b,*]

[a] Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Daejeon 305-806, Republic of Korea
[b] University of Science and Technology (UST), 113 Gwahangno, Yuseong-gu, Daejeon 305-806, Republic of Korea

ARTICLE INFO

Article history:
Received 13 March 2012
Accepted 6 May 2012
Available online 11 May 2012

Keywords:
Conjoined gene
Bioinformatics

ABSTRACT

A conjoined gene is defined as one formed at the time of transcription by combining at least part of one exon from each of two or more distinct genes that lie on the same chromosome, in the same or opposite orientation, which translate independently into different proteins. We comparatively studied the extent of conjoined genes in thirteen genomes by analyzing the public databases of expressed sequence tags and mRNA sequences using a set of computational tools designed to identify conjoined genes on the same DNA strand or opposite DNA strands of the same genomic locus. The CACG database, available at http://cgc.kribb.re.kr/map/, includes a number of conjoined genes (7131—human, 2—chimpanzee, 5—orangutan, 57—chicken, 4—rhesus monkey, 651—cow, 27—dog, 2512—mouse, 263—rat, 1482—zebrafish, 5—horse, 29—sheep, and 8—medaka) and is very effective and easy to use to analyze the evolutionary process of conjoined genes when comparing different species.

1. Introduction

Conjoined genes, also known as read-through transcripts or co-transcribed genes, are defined as genes that give rise to transcripts that combine at least part of one exon from each of two or more distinct genes that lie on the same chromosome [1]. Several studies recently reported that the occurrence of conjoined genes could be advantageous for gene expression regulation or harmful as factors in new disease development during the evolutionary pathway [2–4]. For example, the peroxisome proliferator-activated receptor gamma (*PPARG*) conjoined gene has been intensively studied because it encodes two nuclear receptors (PPARγ1 and PPARγ2) implicated in the regulation of a variety of cellular processes, such as cell cycle control, carcinogenesis, inflammation, atherosclerosis, and mostly adipogenesis [5]. In addition, the SPINT1-C19orf33 conjoined mRNA transcript was found to be transcribed from both SPINT2 and C19orf33 genes in human kidney, prostate, and placenta by Northern blot analysis [6]. In humans, increasing evidence suggests that conjoined genes may play a key role in a range of human diseases [7]. For example, research into the regulation of imprinted genes within the human 15q11–15q13 region have implicated the expression of SNURF–SNRPN and ubiquitin-protein ligase E3A (UBE3A) conjoined genes with a reduction in UBE3A expression which is associated with Prader–Willi and Angelman syndromes [8].

Recently, data regarding intergenic splicing or transcription-induced chimerism in the human genome has been published [9–12]. The number of genes exhibiting occasional intergenic splicing into a single, tandem transcript sequence with the potential of encoding a chimeric protein sequence has been estimated to account for 4–5% of the human genome. This approximation is according to the ENCODE project, whose aim was to identify all functional elements in the human genome, based on 1% of the human genome [13]. Typically, such conjoined transcripts begin at the promoter region of the upstream gene and end at the termination point of the downstream gene. The intergenic region is spliced out of the transcript as an intron so that the resulting fused transcripts possess exons from the two different genes. Various mechanisms appear to be involved in this process and can be classified into two categories; trans-splicing events between pre-mRNAs of distinct genes, and long transcription events across neighboring genes that normally act as independent transcription units.

Bioinformatics approaches with experimental validation have been used to calculate the percentage of conjoined genes based on expressed sequence tags and full-length cDNA sequences. Todd et al. [1] reported the identification of 751 conjoined genes, and experimentally confirmed for the first time the existence of 291 conjoined genes in 16 human tissues. Additionally, a recent transcriptome study using next generation sequencing (NGS) technology identified numerous conjoined genes that were expressed broadly across benign samples and several cancer cell lines without tissue-specific expression [14].

According to some previous studies, conjoined gene transcripts can expand functional protein diversity and play a key role in a

* Corresponding author at: Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Daejeon 305-806, Republic of Korea. Fax: +82 42 879 8139.
E-mail address: hspark@kribb.re.kr (H-S. Park).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.05.005

range of human diseases. In this study we employed an evolutionary approach to address this issue by comparing the genome organization between humans and different species. We performed a comprehensive comparative genomics analysis across thirteen genomes to find intergenic splicing patterns of conjoined genes during genomic evolution. We designed a resource for analyzing and comparing conjoined genes between a wide range of animal genomes and have included genome-wide analyses of conjoined genes from thirteen animal species including human and vertebrates. We utilized the comparative analysis of conjoined genes (CACG) database as the main tool to investigate the evolution of the human genome and diseases by comparing the expression profiles of conjoined genes.

2. Results and discussion

2.1. Searching for conjoined genes from transcript sequences

To characterize the phenomenon of conjoined genes in the vertebrate genomes, we first clustered mRNA and EST transcript sequences onto genome sequences using the BLAT alignment of known genes. Conjoined genes were defined, for the purpose of this study, as a pair of adjacent genes whose genomic regions partly overlap. The EST and mRNA to genome alignment data and the genome sequence assemblies were downloaded from the UCSC database in hg19, mm9, rn4, galGal3, panTro3, rheMac2, oryLat2, bosTau5, oviAri1, canFam2, ponAbe2, equCab2, and danRer7. The EST and mRNA to genome alignments were extracted from the tables all_mrna and all_est (human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka). We performed genome-wide analysis of intergenic splicing events in the genome. From the data, we attempted to map the mRNA sequences to the genome sequences. Minimum length and percent identity of valid alignments were 50 base pairs (bp) and 97%, respectively. To reduce the workload and improve the mapping quality, we first applied the selected sense orientation reliable transcripts. All imperfect alignments were removed. The transcript sequences that were aligned to more than one genomic fragment were discarded as suspected chimeras. All of the putative conjoined genes were also mapped onto the genome. If the Reference Sequence (RefSeq) mRNA sequences overlapped, only the longest was considered. In the results, we extracted the position information of the exon and genome sequences to be matched. Based on this information, the locations of conjoined gene transcripts and exons on each gene were calculated from their position in the genome. Next we searched whether the sequences connecting these two transcripts were canonically spliced, and share at least one splice site with each of the two separate genes. False positive cases arising out of misalignments and alternative splicing of the same loci were removed by manual curation. An algorithm was developed based on the positional comparison of the alignment of the known genes to the mRNA and EST transcript sequences. The algorithm identified all mRNA and EST sequences which aligned to two or more different genes as defined in the UCSC gene database. Finally, our procedure identified 7131 human, 2 chimpanzee, 5 orangutan, 57 chicken, 4 rhesus monkey, 651 cow, 27 dog, 2512 mouse, 263 rat, 1482 zebrafish, 5 horse, 29 sheep, and 8 medaka hits. Furthermore, to illustrate the value of CACG for biological discovery, we analyzed tissue-specificity and cancer versus normal human conjoined genes. We added cDNA libraries to our tissue classification database. Each library source was classified as one of 250 tissue types, and also as abnormal or normal in origin.

2.2. Classification of conjoined genes according to genomic location

Conjoined genes were categorized into three different types according to the following: 1) 5′ splice site at the upstream gene–3′ splice site at downstream gene, 2) 5′ splice site at the upstream gene–3′ splice site at the internal gene, and 3) 5′ splice site at the upstream gene–3′ splice site downstream, in two or more different genes. In these types of conjoined genes, a novel intron is created, spanning the region between the donor of the last intron and the acceptor of the first intron of the upstream and downstream genes, respectively. This abundant conjoined gene splicing pattern usually results in removal of the intergenic region, along with the three prime untranslated region (3′ UTR) and five prime untranslated region (5′ UTR) of the upstream and downstream genes, respectively. Consequently, conjoined gene transcription is terminated at the poly-A signal site of the downstream parent gene; otherwise, even in cases where the signal site is truncated, termination occurs at a poly-A signal site newly created by another event, such as exonization of an intronic sequence.

2.3. Identification of new exons in intergenic regions between two different genes

We further analyzed the creation of novel exons during the formation of conjoined genes which occurs frequently within intergenic regions between two parent genes. First, we divided novel exons into two major groups: 1) TE-containing novel exons (2384 human, 1 sheep, 101 cow, 2 dog, 854 mouse, 91 rat, 7 chicken, 2 rhesus monkey, and 166 zebrafish) and 2) non-TE-containing novel exons according to the fusion of transposable elements in the exon regions. Specifically, we screened TE-containing novel exons for the presence of repetitive sequences using RepeatMasker (http://www.repeatmasker.org). Only the positions of the canonically spliced introns and non-canonical spliced introns that obeyed the GT/AG, GC/AG, AT/AC, GG/AG, CT/AG, GT/CG, GT/TG, AT/AG, GA/AG, GT/GG, TT/AG, and GT/AC rule were considered [15]. Next, we classified the novel exons by three criteria: 1) 3′ UTR, 2) 5′ UTR, and 3) CoDing Sequence (CDS) according to the genome location using genomic mapping data from the UCSC genome browser.

2.4. Comparative analysis of conjoined genes

To investigate the evolutionary landscape of the conjoined genes, we made use of the massive data sets available from comparative genomic studies of thirteen vertebrate genomes. An interesting phenomenon encountered was the gain and loss of conjoined gene states. To explore this observation, we also examined the intergenic splicing states of orthologous conjoined genes in thirteen genomes (human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka). The evolutionary relationships between the conjoined genes in the thirteen analyzed species were assessed by extracting all conjoined genes conserved between species. To analyze the evolutionary impact of conjoined genes among human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka, we compared the human genome to the other genomes. Multiple sequences clustered to each conjoined gene were aligned together using the CLUSTAL W program [16] and the output was computationally and visually inspected to remove alignment errors. To help researchers easily compare conjoined gene data between species, we performed a comprehensive comparative genomic analysis across the thirteen genomes, identifying orthologous conjoined genes, and intergenic splice events between these genomes. Selected species were chosen based on a wide-range of cross-species comparisons with human data, in addition to compiling relatively complete data sets of genomic sequences and abundant transcript sequences. We analyzed gene structure and intergenic splicing via genome-wide studies of thirteen vertebrate genomes to construct the conjoined gene database. This genome-wide database identifies the creation of conjoined genes and conjoined genes loss events during vertebrate evolution.

2.5. User interface

The CACG (comparative analysis of conjoined genes) database is publicly accessible at http://cgc.kribb.re.kr/map/. We stored and managed all the data in MySQL, which is a popular and open source database management system widely used in bioinformatics and biomedical database development. There are various ways for users to access the data stored in the CACG database. The database can be browsed by selecting a specific genome and gene name from the main page. The web interface allows users to access the database content via three different search options. First, users can search genes of interest by using the Human Genome Organization (HUGO) symbol name. In addition, one can use this route to get gene sequences and detailed gene information from the NCBI data bank (Fig. 2A). Second, users can search conjoined genes by clicking one of the genomes listed on the main page (Fig. 2B). The genome browser of CACG will then show annotation features of all the conjoined genes when a particular organism on the multiple genome menu is selected (Fig. 2A). To investigate evolutionary aspects of conjoined genes, we supplemented CACG with a novel visualization interface. The web-based genome browser was executed using Hypertext Preprocessor (PHP) technology, which has the advantage of constructing a clearly defined architecture by separating application logic and presentation. In addition, the CACG database supports a visualization interface that shows the comparative configuration of conjoined genes across multiple species along the whole chromosome scale. The conjoined genes were further aligned to EST/mRNA sequences. A conjoined genes viewer was developed in the database which shows the distribution of conjoined genes in each vertebrate genome. Comparative genomics is a major focus of the CACG web interface, displaying results from new orthologous conjoined genes and species-specific conjoined genes. CACG includes links to comparative genomics information from all views. Users can also search human data for conjoined gene tissue information at the bottom of the gene summary page.

In Fig. 2C, we show one conjoined gene pair (NDUFB8 and SEC31B), which commonly appears in humans and mice. The distance measured between conjoined genes is displayed as a phylogenetic tree (Fig. 2D), enabling users to infer the evolutionary history of the conjoined gene transcripts at a glance. Sense transcripts are depicted in blue and anti-sense transcripts are in red. A small thick segment denotes the exon and a thin line denotes the intron. A short introduction to the web interface and a comprehensive user's guide are available at the CACG website, http://cgc.kribb.re.kr/map/. Moreover, the CACG database incorporates multiple genome and tree visualization tools to facilitate online images of the data. The CACG database contains independent analysis of conjoined genes from EST and mRNA data from thirteen different organisms, including human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka, and thus has many potential future applications, including comparison of conjoined gene patterns between different vertebrate species.

2.6. Summary and future directions

The CACG database is an integrated catalog of conjoined genes which includes bioinformatics analysis data. CACG supports all conjoined genes that are common in a subset of human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka. In addition, it provides a manually curated database that displays the evolutionary features of conjoined genes. The CACG database is very effective and easy to use for comparative analysis and investigation of evolutionary processes involving conjoined genes. Furthermore, the CACG database supports a visualization interface that shows the comparative configuration of conjoined genes across multiple species along the whole chromosome scale. The database is constantly being supplemented with new genome data from a range of available sources. We also plan to supplement this database with conjoined gene information from other mammalian species so that they can be directly compared with human conjoined genes. CACG could potentially be used as the main tool to investigate the evolution of the human genome in relation to diseases by comparing the expression profiles of conjoined genes.

3. Material and methods

We used the combined data of publicly available expressed sequence tags (ESTs), mRNA, and genome alignment from the University of California, Santa Cruz (UCSC) Genome Browser database (http://genome.ucsc.edu; Feb, 2009, release). These alignments were produced by the BLAST-*L*ike *A*lignment *T*ool (BLAT) using ESTs and mRNA databases. The genome data sets (human, chimpanzee, horse, sheep, orangutan, cow, dog, mouse, rat, chicken, rhesus monkey, zebrafish, and medaka) used in our analysis were obtained from the UCSC genome browser database.

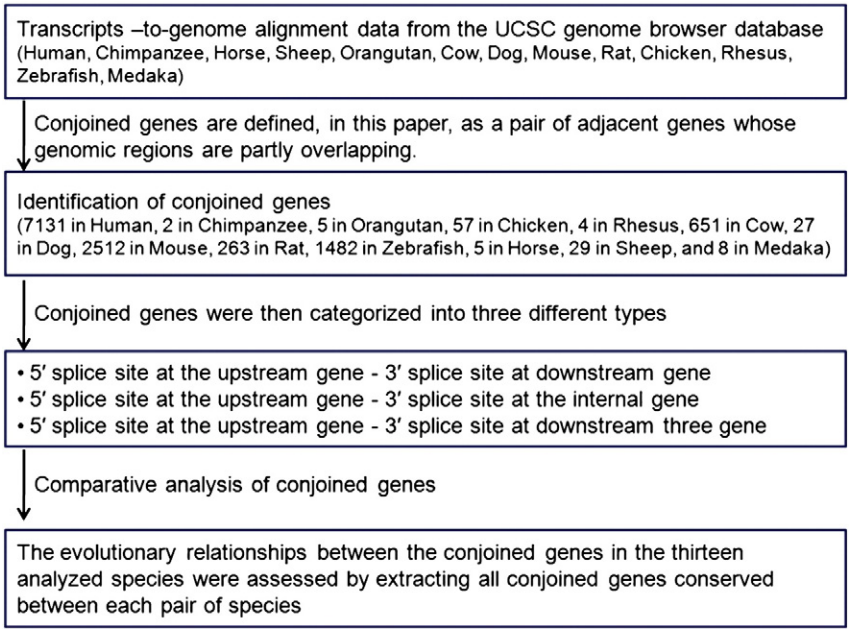

Fig. 1. Flow chart showing the overall procedure used to search for conjoined genes in public sequence databases.

Fig. 2. (A) Visualization of conjoined genes via web retrieval interface. The web interface allows users access to the database contents via three different search options for finding organisms with conjoined genes specified by the user. (B) The users can search interesting overlapping genes by clicking genomes listed on the main page. (C) The easy to use CACG results page is very effective for comparative analysis and investigation of the evolutionary process of conjoined genes. (D) The phylogenic tree calculated by the number of conjoined gene transcripts between each two-genome pair.

3.1. Database design

To build the CACG database, we developed an automatic pipeline consisting of three main steps: (i) collection of conjoined genes from thirteen vertebrate genomes, (ii) classification of splicing patterns of conjoined genes according to genomic location, and (iii) comparative analysis of conjoined genes (Fig. 1).

Acknowledgments

This study was supported by a grant from the Ministry of Education, Science, and Technology (2009-0084206).

References

[1] T. Prakash, V.K. Sharma, N. Adati, R. Ozawa, N. Kumar, Y. Nishida, T. Fujikake, T. Takeda, T.D. Taylor, Expression of conjoined genes: another mechanism for gene regulation in eukaryotes, PLoS One 5 (2010) e13284.
[2] M. Kato, S. Khan, N. Gonzalez, B.P. O'Neill, K.J. McDonald, B.J. Cooper, N.Z. Angel, D.N. Hart, Hodgkin's lymphoma cell lines express a fusion protein encoded by intergenically spliced mRNA for the multilectin receptor DEC-205 (CD205) and a novel C-type lectin receptor DCL-1, J. Biol. Chem. 278 (2003) 34035–34041.
[3] K. Wang, G. Ubriaco, L.C. Sutherland, RBM6–RBM5 transcription induced chimeras are differentially expressed in tumours, BMC Genomics 8 (2007) 348–360.
[4] P.E. Kowalski, J.D. Freeman, D.L. Mager, Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes, Genomics 57 (1999) 371–379.
[5] M. Roux, H. Levéziel, V. Amarger, Cotranscription and intergenic splicing of the PPARG and TSEN2 genes in cattle, BMC Genomics 7 (2006) 71–82.
[6] S. Naganuma, H. Itoh, S. Uchiyama, H. Tanaka, K. Nagaike, S. Miyata, S. Uchinokura, Y. Nuki, Y. Akiyama, K. Chijiiwa, H. Kataoka, Characterization of transcripts generated from mouse hepatocyte growth factor activator inhibitor type 2 (HAI-2) and HAI-2-related small peptide (H2RSP) genes: chimeric mRNA transcribed from both HAI-2 and H2RSP genes is detected in human but not in mouse, Biochem. Biophys. Res. Commun. 302 (2003) 345–353.
[7] M. Kumar, G.G. Carmichael, Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes, Microbiol. Mol. Biol. Rev. 62 (1998) 1415–1434.
[8] M. Runte, A. Hüttenhofer, S. Gross, M. Kiefmann, B. Horsthemke, K. Buiting, The IC-SNURF–SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A, Hum. Mol. Genet. 10 (2001) 2687–2700.
[9] P. Akiva, A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, et al., Transcription-mediated gene fusion in the human genome, Genome Res. 16 (2006) 30–36.
[10] G. Parra, A. Reymond, N. Dabbouseh, E.T. Dermitzakis, R. Castelo, T.M. Thomson, et al., Tandem chimerism as a means to increase protein complexity in the human genome, Genome Res. 16 (2006) 37–44.
[11] F. Denoeud, P. Kapranov, C. Ucla, A. Frankish, R. Castelo, J. Drenkow, J. Lagarde, T. Alioto, C. Manzano, J. Chrast, S. Dike, C. Wyss, C.N. Henrichsen, N. Holroyd, M.C. Dickson, R. Taylor, Z. Hance, S. Foissac, R.M. Myers, J. Rogers, T. Hubbard, J. Harrow, R. Guigó, T.R. Gingeras, S.E. Antonarakis, A. Reymond, Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions, Genome Res. 17 (2007) 746–759.
[12] A. Siepel, M. Diekhans, B. Brejová, L. Langton, M. Stevens, C.L. Comstock, Targeted discovery of novel human exons by comparative genomics, Genome Res. 17 (2007) 1763–1773.
[13] ENCODE Project Consortium, The ENCODE (ENCyclopedia of DNA Elements) Project, Science 306 (2004) 636–640.
[14] C.A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, A.M. Chinnaiyan, Transcriptome sequencing to detect gene fusions in cancer, Nature 458 (2009) 97–101.
[15] T.A. Thanaraj, F. Clark, Human GC–AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions, Nucleic Acids Res. 29 (2001) 2581–2593.
[16] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.